EXHIBIT G

Refutation of Data Colada Claims about the 2012 PNAS Paper

Among Data Colada's four-part series disparaging of my work is their June 17, 2023, blog post relating to Study 1 in the 2012 paper "Signing at the beginning makes ethics salient and decreases dishonest self-reports in comparison to signing at the end" (published in the *Proceedings of the National Academy of Sciences*, in short form "2012 PNAS paper"). Data Colada alleged that the study relies on data manipulation. An HBS report concluded much the same thing. **They are both wrong**.

The Allegation

Data Colada's critique of Study 1 is grounded in three inferences they draw from the data posted to the Open Science Framework (OSF).

1. Data Colada identifies **eight rows of data that they consider suspicious**. Two of these rows have duplicate IDs and six have IDs that are out-of-sequence.
2. Data Colada claims that these eight suspicious rows have **extreme values, showing a huge effect.**
3. Data Colada claims that the eight suspicious rows were a consequence of data tampering by me. They use Excel's **calcChain feature as evidence for their claim.**

All three inferences are wrong, as I establish in this post.

Data Colada cherry-picked the data it chose to include in its analysis.

It is true that the data in this study includes rows with duplicate IDs, as well as rows with out-of-order sequences. Below, I will describe how this is likely to have happened. But first, it is important to note that Data Colada cherry-picked the observations to include in its analysis—using observations that supported their claim that I manipulated data, while omitting observations that weakened their claim.

Specifically:

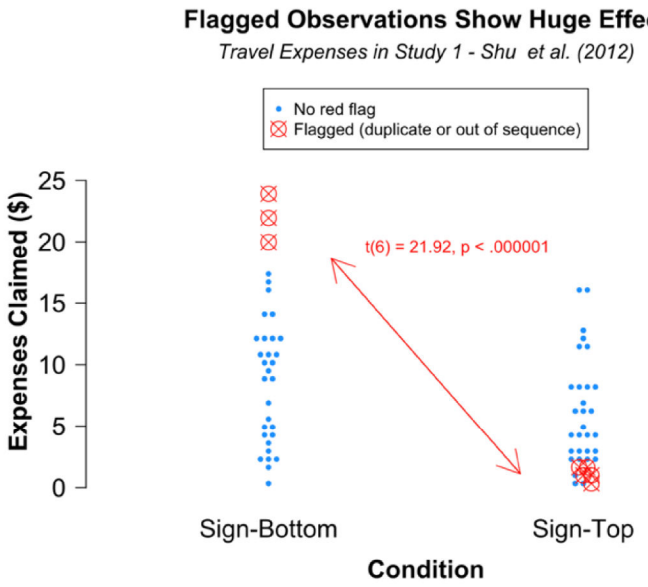
- When it comes to duplicate IDs, Data Colada included rows 52 and 53 (ID 49), but they **omitted** rows 5 and 75, which also have duplicate IDs (ID 13).
- When it comes to out-of-sequence observations, Data Colada included six observations, but they did three very curious things here:
  - They included two observations that are in fact **not** out-of-sequence—namely, Row 69 (ID 101) and Row 70 (ID 7).
  - They mentioned Row 33 (ID 64) as being out-of-sequence, but they **omitted** it from their analysis.
  - Although Row 73 (ID 5) is also out-of-sequence, they **failed to mention** it, and they **omitted** it from their analysis.

It's troubling that Data Colada cherry-picked data in this manner and without explanation. I've included additional details about these errors in **Appendix 1**.

In addition, Data Colada excluded one of the three conditions in the study. This exclusion of the third condition, combined with the cherry-picking of data, enabled Data Colada to create an enormously misleading impression.

The published study actually had three experimental conditions. Data Colada only examined conditions 1 and 2, entirely omitting condition 0 (the control condition). By omitting this control condition and cherry-picking observations within conditions 1 and 2, Data Colada was able to make the argument that the suspicious rows contained extreme values that drove huge effects.

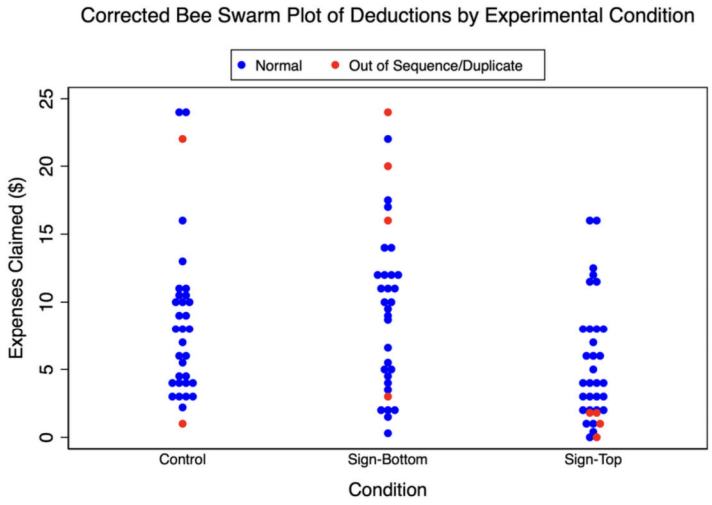
According to their analysis, when running a t-test considering the eight observations they flagged in conditions 1 and 2 of the study, the p-value is highly significant [$t(6) = 21.92$, $p < .000001$]. Here is the compelling figure they used to underscore their conclusion:



I re-ran the analysis across all conditions, using all of the duplicate and out-of-sequence observations while omitting the two observations that actually were not out of sequence. This leaves 10 observations, and the test results for these observations are far less dramatic. In fact, **these ten do not even reach statistical significance** [$F(2,9) = 3.28$, $p = .099$].

In other words, using Data Colada's very own "rules" for flagging suspicious data, the "huge effect" disappears when the observations are no longer cherry-picked.

Here is the correct figure:



As the corrected bee swarm plot shows, the Data Colada narrative collapses when including the third condition and the correct data points for conditions 1 and 2.

Furthermore, when I excluded all 10 of the flagged observations due to their "suspicious" nature, and then re-ran the analysis using the remaining observations, the findings of the original study still hold.

This finding speaks directly to motive. My accusers think I manipulated data to drive results. But this analysis demonstrates that even without all the observations that, according to Data Colada's own rules, should be considered "suspicious," the findings of the original study still hold.

Because this point is of such critical importance, I've included additional analyses in Appendix 2. These analyses also re-run the study results using all three dependent variables (DVs) used in the original study. This is another critical point: The original study had three DVs. Data Colada inexplicably only analyzes one of these three DVs in their blog post, perhaps because an analysis of the other two DVs fails to support their allegations at all. This is another example of Data Colada making a deliberate choice about what to share with readers, and what not to share. Their post offers no explanation for this, nor do they warn readers that they report results from only one of the three DVs, and the other two show no such pattern.

Data Colada completely misrepresented Excel's calcChain feature in order to buttress their claim that I tampered with data

I have recently learned a lot about Excel's calcChain feature. I carefully reviewed the documentation Microsoft provides about this feature, and I ran hands-on tests to figure out how it works.

Data Colada claimed they were able to use Excel's calcChain feature to tell "whether a cell (or row) containing a formula has been moved, and where it has been moved to." Data

Colada also claimed an analyst could "use calcChain to go back and see what this spreadsheet may have looked like… before it was tampered with."

These statements are both misleading and irresponsible. No Microsoft documentation supports using calcChain in this way or indicates calcChain was designed to work this way. Quite the contrary, Microsoft is clear in stating calcChain's actual purpose ("indicates the order in which the cells were last calculated") and even in alerting readers that Excel can change calcChain on its own as it optimizes calculation speed. As I've learned, the way Data Colada purported to use calcChain is not the way calcChain should ever be used, because **calcChain just doesn't do what Data Colada says it does.**

Three examples show the rows in calcChain can get out-of-sequence through benign actions, without any manipulation of data whatsoever, as well as how it is possible to manipulate data and not have it show up in calcChain. You can refer to **Appendix 3** for a detailed description of these examples, including instructions on how to reproduce the results:

- **Example 1:** Excel's "fill cursor" puts calcChain out-of-sequence. Most Excel users have used Excel's "fill cursor," a plus sign that, when dragged, extends a formula into adjacent cells. This is a simple, routine Excel procedure. In **Appendix 3a**, I demonstrate how using the "fill cursor" – without any manipulation of order or anything else – can result in a calcChain that is out-of-sequence.
- **Example 2:** Some Excel move commands create an out-of-sequence calcChain, but others don't. In **Appendix 3b**, I compare two features that users often invoke interchangeably to achieve the same effect: Copy-Paste and Drag-and-Move. In the latter case, calcChain tracks the change. In the former case, calcChain does not. In other words, some of these moves are recorded; others are not.
- **Example 3:** Sorting can reorder a calcChain. When a user uses the sorting function on an Excel dataset in a way that rearranges entries with formulas, the calcChain tracks the change in a way that is unpredictable but appears to depend on the relationship between the cells with formulas. In **Appendix 3c**, I provide an example.
- **Example 4:** Excel itself can reorder a calcChain. Indeed, Excel can reorder a calcChain without a user moving anything, based on Excel's assessment of what calculation sequence is fastest. In **Appendix 3d**, I provide an example.

I have also discovered that Data Colada cherry-picked data from the calcChain file as well. In their blog post, Data Colada notes that the 6 suspicious observations they flagged have out-of-order appearance in the calcChain file, which they present as support for their finding of data tampering. But they fail to mention that calcChain file actually shows 69 (!) out-of-sequence observations, not just the 6 they discuss. (This underscores how easy it is for a user to create out-of-sequence rows in calcChain through benign actions.) This further raises questions about what Data Colada chose to share with its readers and what it chose to hide.

In sum, calcChain doesn't work the way Data Colada says it does, and is plainly unsuited to the purpose for which Data Colada invokes it.

My Best Hypothesis as to What Likely Happened In This Study

Before I try to evaluate what caused data glitches in this study, let me remind everyone that the passage of time makes it hard to figure it out definitively. The data was collected in July 2010, more than 13 years ago, in the behavioral lab at UNC. Data was collected by a lab manager with the help of one or more research assistants (RAs). The original data from the study no longer exists, as the study was conducted on paper. This is not unusual: The norm in the field is to discard any paper records after three years, unless required by an applicable sponsor. (HBS uses a Harvard policy that sets out a 7-year rule, endorsing the principle that long-ago work is especially difficult to investigate.)

I recall the behavioral lab at UNC consisting of two small rooms where 4 or 5 participants completed a study at the same time. In this study, each participant received an index card with a number on it, which was used to represent their ID. Each person was supposed to

receive a unique ID number in increasing value, so the first participant would have received an index card with an ID number of 1, the second participant a 2, and so on.

Duplicate IDs can occur in a paper-based study

Back in 2010, providing cards to participants was a manual, paper process. So you can see how mistakes could creep in. In particular, a card could easily be reused in two different sessions. Or the same number could accidentally be written on two different cards. In this case, we're talking about two duplicate ID values – sloppy and unfortunate, yes, but perhaps not terribly surprising given the paper-based system that was used.

Out-of-sequence data can occur in a paper-based study

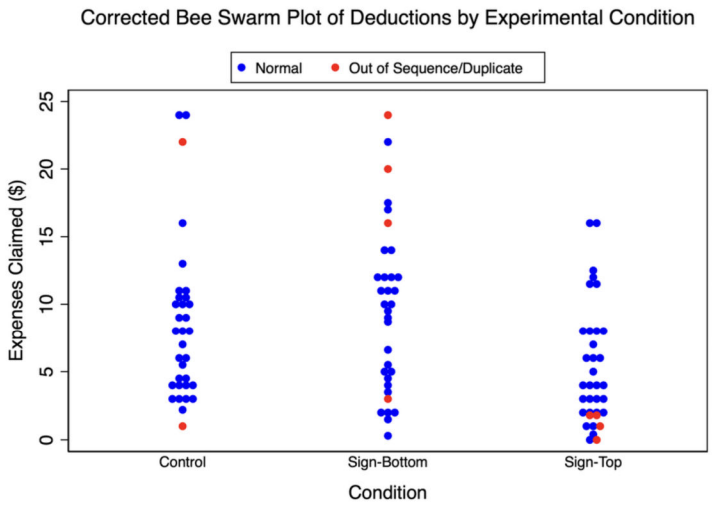
One of the biggest assumptions Data Colada made in their blog post was that the data were electronically sorted twice: First by treatment, and then by participant ID. This assumption is almost certainly incorrect. It is far more likely that the data were never electronically sorted. This is a critical point.

Instead, here is what almost certainly happened: The RAs conducting the study simply stacked the paper copies by condition, and then manually entered the data in the order in which the papers were stacked. The sequence of paper entries tended to be in ascending sequence; this makes sense, since it was probably the same sequence in which the participants performed the tasks. On the other hand, it is also plausible that some of the papers got mis-ordered. After all, no one had any particular reason to care about sort order; the findings and statistical analysis depended on the data points themselves, not their sequence in a file.

Of course, there will still be skeptics who ask – "if that's the case, then why were most of the papers in sequence, and just a few out-of-sequence?" The honest answer is, I do not know. I only know that working with paper creates the potential for mis-ordering to occur.

Finally, there will also be skeptics who ask – "why is it that the out-of-sequence observations just happened to have such extreme values? Shouldn't that be considered suspicious?"

My response to that is simple. Look at the chart below and and try to find the pattern. Fact is, some out-of-sequence/duplicate data points are high, some are low, and some are in the middle. Some extreme values are out-of-sequence or duplicate (red), and others are not (blue). When running studies with human subjects, some of the observations end up having extreme values. This is just how the data came out in this particular study.



Finally, recall the lack of motive for the supposed manipulation: *If you re-run the entire study excluding all of the red observations (the ones that should be considered "suspicious" using Data Colada's lens), the findings of the study still hold.* **Why would I manipulate data, if not to change the results of a study?**

Appendix 1

Appendix 3d

Appendix 2

Appendix 3a

Appendix 3b

Appendix 3c

Additional Analyses on Duplicate IDs and calcChain

About | Innocence | Policy Injustice | A Broken Process | Why A Lawsuit | Case Documents | Index

More About Me

Contact